

Python for Data Science

6. Python Libraries for Data Science



Contents

6.1 Numpy

6.2 Scipy

6.3 Pandas

6.4 Matplotlib

6.5 Tensorflow

Numpy

Stands for numerical Python.

General purpose array-processing package.

Numpy is the fundamental package for numerical computation with Python, it contains a powerful N-dimensional array object.

Features

Provides fast precompiled functions for numerical routines.

Array oriented computing for better efficiency.

Supports object-oriented approach.

Compact & faster computations with vectorization.

Applications

Extensively used in data analysis.

Creating powerful N-dimensional array.

Forms the base of other libraries like Scipy, Scikit-learn.

Replacement of Matlab when used with Scipy, Matplotlib.

Scipy

Stands for scientific Python.

Used for scientific & technical computations.

It provides many user-friendly and efficient routines for scientific computation, it extends Numpy.

Features

A collection of mathematical algorithms and scientific functions built on the Numpy extension of Python.

High-level commands and classes for manipulating and visualizing data.

Multi-dimensional image processing.

Includes functions for computing integrals numerically, solving differential equations, optimization, ..

Applications

Multi-dimensional image operations.

Solving differential equations & Fourier transform.

Optimization algorithms.

Linear algebra.

Pandas

Stands for Python data analysis library.

Used for data analysis and cleaning.

It provides fast, flexible, and expensive data structures designed to work with structured data easily and intuitively.

Features

Eloquent syntax & rich functionality.

Enables you to run a function across a series of data.

High level abstraction.

Contains high-level data structures & manipulation tools.

Applications

General data wrangling.

ETL jobs & data storage.

Used in a wide variety of academic and commercial domains, including statistics.

Time series-specific functionality.

Matplotlib

Plotting library for Python.

Used for data visualization.

It provides an object-oriented API for embedding plots into applications.

Features

As usable as Matlab with an advantage of being free and open-source.

Supports dozens of backends & output types.

Pandas itself can be used as wrappers around Matplotlib's API.

Smaller memory consumption & better runtime behavior.

Applications

Correlation analysis of variables.

Visualize 95% confidence intervals of the models.

Outlier detection.

Visualizing distributions to gain instant insights.

Tensorflow

Library for high performance numerical computation.

Used across many scientific domains.

Data in Tensorflow are represented as tensors, which are multidimensional arrays.

Features

Better computational graph visualizations.

In neural machine translation, reduces error by 50-60%.

Parallel computing to execute complex models.

Seamless library management.

Applications

Speech/Image recognition.

Text based applications.

Time series.

Video detection.